DOCUMENT RESUME

ED 236 169 TM 830 656

AUTHOR Holmes, Susan E.; Doody-Bogan, Evelyn N.
TITLE An Empirical Study of Vertical Equating Methods Using the Three-Parameter Logistic Model.
PUB DATE 12 Apr 83
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Equated Scores; Estimation (Mathematics); *Latent Trait Theory; *Mathematical Models; Reliability; Research Design; Standardized Tests; Statistical Analysis
IDENTIFIERS Comprehensive Tests of Basic Skills; Prescriptive Reading Inventory; *Three Parameter Model; *Vertical Equating

ABSTRACT

The accuracy of trait estimates obtained from three vertical equating methods was examined. The procedures studied included two anchor test designs and a single-group design. Data from two content areas and two grade combinations were studied. A three-parameter logistic model was used to perform the equatings. The results obtained were used to produce pairs of true score estimates for a cross-validation sample, one estimate for each of the equated tests. Comparisons among pairs of estimates indicated that the single-group method consistently produced the most accurate results, although the advantage was small. A differential effect due to content area was reported for the two anchor test methods. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236169

An Empirical Study of Vertical Equating Methods Using the Three-Parameter Logistic Model

by

Susan E. Holmes
Occupational Assessments
and
Evelyn N. Doody-Bogan
Research
CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. E. Holmes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting
of the National Council on Measurement in Education
held in Montreal, April 12, 1983.

TM 830 658

Abstract

The accuracy of trait estimates obtained from three vertical equating methods was examined. The procedures studied included two anchor test designs and a single-group design. Data from two content areas and two grade combinations were studied. A three-parameter logistic model was used to perform the equatings. The results obtained were used to produce pairs of true score estimates for a cross-validation sample, one estimate for each of the equated tests. Comparisons among pairs of estimates indicated that the single-group method consistently produced the most accurate results, although the advantage was small. A differential effect due to content area was reported for the two anchor test methods.

An Empirical Study of Vertical Equating Methods Using the Three-Parameter Logistic Model[1]

Introduction

In most large-scale testing programs, different forms of the same test are used to measure a trait or ability at different educational levels. Since comparisons across educational levels are highly desirable in these programs, the scores from such test forms must somehow be defined in terms of a common scale. The process of determining a scale for test forms that measure the same trait but that differ in difficulty is called "vertical equating."

A wide variety of conventional equating methods can be found in the research literature (Angoff, 1971). When applied to different tests that measure the same trait at the same educational level, these methods have been found to be quite satisfactory. However, less favorable results have been reported when these methods have been applied to vertical equating situations which involve different tests that measure the same trait but at different educational levels (Lord, 1975, 1977). For this reason, a considerable amount of attention is being directed toward vertical equating methods based on an innovative approach to educational measurement called item response theory (IRT). In fact, it has been argued that, strictly speaking, only equating methods based on this approach are appropriate when equating tests which differ in difficulty (Lord, 1975).

[1]The authors wish to acknowledge George Burket, D. Ross Green, and Wendy Yen for their respective contributions to the research reported here.

Item reponse theory assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's trait level (see Lord, 1980, Chapter 2, for details). The model chosen for this study was the three-parameter logistic model. Here, an examinee's trait level is represented by θ. The probability of a correct response to item i, $P(\theta)$, then becomes:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} , \qquad (1)$$

where a_i, b_i, and c_i are three item parameters that describe the item. b_i is item difficulty, a_i is item discrimination, and c_i is the pseudo-chance-success level.

One of the most important advantages of item response theory is that once the item parameters described above have been estimated for a particular set of items, any subset of those items can be used to estimate an examinee's trait level. In addition, all trait estimates produced from these items will be on the same scale, provided that the data fit the model. It is this important measurement outcome that makes the applications of item response theory to test equating possible.

All equating methods have two components, a design for data collection and a statistical model for analyzing data. Three methods were selected for this study. Two of the methods used an external anchor-test design (Angoff, 1971). The third method was a single-group design and served as a criterion against which to compare the other two. All three of the methods involved groups of examinees at different educational levels who had responded to two of three tests. A discussion of each of the three methods follows next.

Method 1. Vertical equating Method 1 can most easily be understood by examining Figure 1. The item response data from a sample of examinees (Sample I) at one educational level are combined and analyzed simultaneously with those of a second sample (Sample II) at a different educational level. Examinees in the first sample responded to Test 1 and the Anchor Test, while those in the second sample responded to Test 2 and the Anchor Test. The Anchor Test is used to measure and take into account differences in the trait levels of the two samples. Because the three tests are analyzed simultaneously, their item parameter estimates are on a common scale thus vertically equating Tests 1 and 2. This equating method has been called "concurrent calibration" by Petersen, Cook, and Stocking (1981). It will be referred to here as "concurrent calibration with an anchor test".

Method 2. Vertical equating Method 2 is applied to the same data collection design as the one described for Method 1. In Method 2, however, each test and anchor test pair are analyzed separately and then equated using the linear relationship between the pairs of anchor test item difficulty estimates (Lord & Novick, 1968). This method of equating, referred to here as "anchor test b-value equating," is similar to the "partial pre-calibration with equated b's" method studied by Petersen, Cook, and Stocking (1981).

Method 3. Vertical equating Method 3 is a second concurrent calibration method. In this case, both of the tests to be equated are taken by the same group of examinees. In this "single-group" design (Lord, 1975), the item response data for both tests are analyzed simultaneously. The item parameter estimates produced by this analysis are automatically on the

same scale, thus vertically equating the two tests. This method is referred to here as "concurrent calibration without an anchor test", and it served as the criterion against which the other methods were compared.

The three-parameter model was used to perform the vertical equatings. First the item parameter estimates of the two tests being equated were placed on the same scale using one of the methods above. Next, these re-scaled item parameter estimates were used to produce two trait estimates for each examinee in a cross-validation sample. One estimate was produced for each of the tests that had been equated. Finally, pairs of trait estimates were compared to determine the accuracy of these vertical equating methods.

Method

Data

Each of the three vertical equating methods examined were studied using data from two different grade combinations: Grades 4 and 5, and Grades 8 and 9. Eight samples were identified in all, four at each of the two grade combinations. Three of the four samples were used to equate Test 1 and Test 2. The fourth sample was used for cross-validation purposes.

The ideal data configuration for this study would be one in which two samples, one at each of two different educational levels, had been administered three tests: Test 1, an Anchor Test, and Test 2. The three vertical equating methods in Figure 1 could then be studied by analyzing the pair of tests appropriate for that method. Such a data configuration is not generally available, however. The three vertical equating methods in Figure 1 were therefore emulated using data from the fall 1980 and spring 1981 standardization of the Comprehensive Tests of Basic Skills (CTBS-U), Form U (CTB/McGraw-Hill, 1981) and from an equating study of the Prescriptive Reading Inventory (PRI), (CTB/McGraw-Hill, 1976) conducted concurrently with the standardization. A nationally representative sample was selected for the standardization. School districts included in this sample were given the option of participating in the PRI equating study.

As noted above, eight samples were identified. Samples I, II, V, and VI (n=1000) were randomly selected from the total available PRI equating data pool. These data were used for equating purposes in Methods 1 and 2. Samples III and VII (n=1000) were also randomly selected from the standardization data pool. These data were used for equating purposes in Method 3. A constant sample size of 1000 cases was maintained for all methods since sample size can affect item parameter estimates. The remaining available cases in the

standardization pool became Samples IV (n=1792) and VIII (n=3935), respectively. These samples were used for cross-validation purposes for all methods.

Instrumentation

Two instruments were used in the study. The first was CTBS-U, an achievement battery containing tests in reading, language, mathematics, reference skills, science, and social studies. The battery consists of nine levels which span the grades K-12. The levels are lettered A through H and J. For the purposes of this research, items from only the vocabulary and reading comprehension tests at levels F, G, H, and J were studied.

In addition to the "shelf" CTBS-U tests just described, two "linkage booklets" from the CTBS-U standardization study were also included in this study. Linkage booklets contained some items from each of two adjacent test levels and were created specifically to facilitate the CTBS-U interlevel equating. The two linkage booklets used in this study were CTBS-F/G and CTBS-H/J.

The second instrument used in the study was the Prescriptive Reading Inventory (PRI), a criterion-referenced test of reading and language arts (CTB/McGraw-Hill, 1976). This battery consists of five levels that span grades K-9+. The levels are lettered A through E. Two of these levels, Levels D and E, were included in this study.

A total of eight tests were therefore studied. They are presented in Table 2 along with the number of items in each.

Procedure

Twelve vertical equatings were performed in all: one for each grade combination (Grades 4/5 or Grades 8/9), content area (vocabulary or comprehension), and vertical equating method (Method 1, 2, or 3). In each case, Test 1 and Test 2 were equated by placing their item parameter estimates on a common scale.

In Method 1, Test 1 and Test 2 were equated by analyzing them and a common anchor test simultaneously. The samples used in this analysis were from two different educational levels. Data for the common anchor items were pooled across samples. The "not reached" option in LOGIST was employed for items unique to only one test (Wood, Wingersky, & Lord, 1976.) The item parameter estimates produced by such an analysis are automatically on a common scale, thus vertically equating Tests 1 and 2.

In Method 2, the same samples were used as in Method 1. In this case, however, the data from each sample were analyzed separately. The linear relationship between pairs of item difficulty estimates (b_i's) for the Anchor Test items (Lord & Novick, 1968) was used to place the item parameter estimates from Test 1 on the same scale as Test 2. The transformations applied are given by the following expressions:

$$b_1^* = \left(\frac{SD_{b2}}{SD_{b1}}\right) b_1 + \left[\bar{b}_2 - \left(\frac{SD_{b2}}{SD_{b1}}\right) \bar{b}_1 \right] ,$$

$$a_1^* = a_1 / \left(\frac{SD_{b2}}{SD_{b1}} \right) , \text{ and}$$

$$c_1^* = c_1 .$$

Here a_1^*, b_1^*, and c_1^* are the transformed item parameter estimates for Test 1. $\bar{b}_1$ and $\bar{b}_2$ are the mean item difficulty estimates for the Anchor Test items when those items were analyzed with Tests 1 and 2, respectively. SD_{b1} and SD_{b2} are their respective standard deviations.

In Method 3, both Test 1 and Test 2 were taken by the same group of examinees. The item parameter estimates for these items were placed on the same scale by simultaneously analyzing the data from both samples. This equating method is identical to the one used with Method 1 except that no anchor test was included since the same group of examinees had taken both Test 1 and Test 2.

After each equating had been performed, the item parameter estimates for Test 1 and for Test 2 were applied to the item response data of the cross-validation samples. Two trait estimates were produced for each examinee in both samples: one estimate for Test 1 and one estimate for Test 2. This was possible because examinees in the cross-validation sample had actually responded to both tests.

Since maximum likelihood estimation procedures were used, trait estimates could not be produced for examinees who earned either zero or perfect scores on a test (see Lord, 1980, pp. 182-183). Deleting such examinees from the comparisons could, however, bias the results obtained. Following a procedure suggested by Lord (1980, p. 183), the following transformation was therefore applied to all trait estimates:

$$\hat{\zeta} = \zeta(\hat{\theta}) = \frac{1}{n_1 + n_2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} P_i(\hat{\theta}) \quad , \qquad (2)$$

where $P_i(\hat{\theta})$ is the probability of a correct response to item i as given by Equation 1, and n_1 and n_2 are the number of items in Test 1 and Test 2, respectively. Examinees with zero or perfect scores were assigned $\hat{\zeta}$-values of $\overline{c}$ or 1.0, respectively.

Using Equation 2, pairs of proportion-correct true scores, here referred to simply as "true score estimates," were produced for each examinee in the cross-validation sample. The theoretical range of these values is from $\overline{c}$ to 1.0.

The following comparisons between pairs of true score estimates were then made for each of the three methods. The Pearson product-moment correlation between pairs of estimates was computed and a root mean squared difference (RMSD). The latter was scaled by the variances of the two true score estimates being compared and is given by the following expression:

$$\frac{\text{RMSD}}{S_{\hat{\zeta}}} = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{\zeta}_{1j} - \hat{\zeta}_{2j})^2}}{\sqrt{(S^2_{\hat{\zeta}_1} + S^2_{\hat{\zeta}_2})/2}} \quad . \qquad (3)$$

Here $\hat{\zeta}_{1j}$ and $\hat{\zeta}_{2j}$ are two true score estimates of the same trait for examinee j (j=1,n examinees), and $S^2_{\hat{\zeta}1}$ and $S^2_{\hat{\zeta}2}$ are the variances of the two estimates. RMSD values were rescaled in the manner to facilitate comparisons across designs.

Results

Test Summary Information

Raw score summary statistics for the tests studied can be found in Tables 3 and 4 by test pair. The following information is reported there for Test 1, the Anchor Test, and Test 2: the number of items in each test, the raw score mean, and the standard deviation. Also reported is the Pearson product-moment correlation between test pairs. Table 3 contains the information for the vocabulary tests, while Table 4 contains the information for the comprehension tests.

Insert Tables 3 and 4
about here

Equating Methods

The results of the 12 equatings performed in the study are presented in Tables 5 and 6 for the vocabulary and comprehension tests, respectively.

Insert Tables 5 and 6
about here

Test 1 and Test 2 were equated by placing their item parameter estimates on a common scale using one of the three equating methods. The resulting item parameter estimates were then applied to the data from a cross-validation sample to produce pairs of true score estimates. The average true score estimate for Test 1 and Test 2($\bar{\hat{\zeta}}$) and the standard deviation of that estimate ($SD_{\hat{\zeta}}$) are presented by content area and grade level combination. Also reported there is the ratio of the standard deviations of the true score estimates, the scaled root mean squared difference between pairs of true score estimates and the Pearson product-moment correlation between them.

Discussion

Appropriateness of the Data

The data from eight samples of examinees were used in this research, four samples at each of two grade combinations. Although the data were not specifically collected for this research, the raw score means and standard deviations reported in Tables 3 and 4 suggest that, with only one exception, the data were appropriate for a study of vertical equating. For example, in Table 3 the raw score means for Test 1 (Grade 4.8) were 15.54 and 15.39 for Samples III and IV, respectively. These values are larger than the raw score mean of 13.90 reported for Sample I (Grade 4.1) for the same test. Similarly, the raw score means for Test 2 (Grade 4.8) were 12.26 and 12.07 for Samples III and IV, respectively. These values are smaller than the raw score mean of 13.01 for Sample II (Grade 5.1) for the same test. In addition, in three of the four situations studied, the raw score mean for Test 1 was considerably larger than the raw score mean for Test 2 for the same sample suggesting that a verical equating situation existed. This was not the case for the comprehension tests at the lower grade combination (Table 4). Here the raw score means for Test 1 and Test 2 were 13.61 and 13.45, respectively, for Sample III and 13.38 and 13.14, respectively, for Sample IV. Tests that are so similar in difficulty represent a hozizontal equating situation.

Vertical Equating

As would be expected, Method 3, the second concurrent calibration method, yielded the most accurate vertical equating of Test 1 and Test 2, regardless of grade combination or content area. In general, this method produced the largest correlations between true score estimates and the smallest scaled root mean squared differences, regardless of content areas or grade combination.

However, as can be seen in Tables 5 and 6, the advantage of this method over the others was small in some cases.

Although Method 3 produced the most accurate equating results in every situation, all three vertical equating methods yielded more accurate equating results at the lower grade combination studied, regardless of content area, and for the vocabulary tests, regardless of grade combination or the fact that the comprehension tests for the lower grade combination were closer to a horizontal than to a vertical equating situation. Larger correlations between true score estimates and smaller scaled root mean squared differences were observed for these situations. The raw score correlations reported in Tables 3 and 4 also reflect these results. All three methods produced larger correlations between Test 1 and Test 2 for the lower grade combination and for the vocabulary tests. Further, the correlations between Test 1 and the Anchor Test and between Test 2 and the Anchor Test were generally larger for the lower grade combination and for the vocabulary tests. This finding underscores the importance of selecting appropriate anchor tests when performing vertical equating using an anchor-test design (Lord, 1975, p. 2-3).

When comparing only Methods 1 and 2, a differential effect due to content area was observed. Method 1 produced consistently smaller root mean squared differences when used to equate the vocabulary tests, while Method 2 produced smaller values for the comprehension equatings, regardless of grade combination. Some slight advantage was noted for Method 1 with respect to the standard deviations of the true score estimates it produced.

Differential results due to content area were also observed in a comparison of several IRT equating methods reported by Petersen, Cook and Stocking (1981, p. 33). In that study, a procedure similar to Method 2, anchor b-value equating, was found to be more satisfactory than Method 1 for verbal data, while the converse was true for the mathematics data studied by these authors. IRT methods were, however, reported to be superior to the conventional methods studied.

In general, the results reported for the three vertical equating methods studied were quite similar. This may have been due to the samples used in the study. First, the cross-validation samples were randomly selected. Larger differences between equating methods might have been observed if non-random cross-validation samples had been employed. Second, the grade level of the cross-validation samples was different from that of the samples used in the equatings. Samples at grade levels 4.1 and 5.1 (or 8.1 and 9.1) were used to perform the equatings in Methods 1 and 2. The grade level of the cross-validation sample, however, was 4.8 (or 8.8). Samples at the same grade levels as the cross-validation samples (4.8 or 8.8) were also used to perform the equatings in Method 3. Larger differences between vertical equating methods might have been observed if the cross-validation sample had included the same grade levels as the samples used to perform the equatings. Such a sample might include more examinees at the extremes of the ability distribution. Such a study is currently underway.

References

Angoff, W.H. Scales, norms and equivalent scores. In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Comprehensive Tests of Basic Skills, Form U. Monterey, CA: CTB/McGraw-Hill, 1981.

Lord, F.M. A survey of equating methods based on item characteristic curve theory. (ETS RB 15-13). Princeton, N.J.: Educational Testing Service, 1975.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Company, 1968.

Petersen, N.S., Cook, L.L., & Stocking, M.L. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles 1981.

Prescriptive Reading Inventory. Monterey, CA: CTB/McGraw-Hill, 1976.

Wood, R.L., Wingersky, M.S., & Lord, F.M. A computer program for estimating examinee ability and item characteristic curve parameters. (ETS RB 76-6). Princeton, NJ: Educational Testing Service, 1976.

Yen, W.M. Using simulation results to choose a latent-trait model. Applied Psychological Measurement, 1981, 5, 245-262.

Table 1

The Eight Samples Identified, Their Grade Placement, Number of Cases, and Test Pair Taken

Sample	Grade	n	Test Pair	Purpose
I	4.1	1000	CTBS-F + PRI-D	Equating
II	5.1	1000	CTBS-G + PRI-D	Equating
III	4.8	1000	CTBS-F/G	Equating
IV	4.8	1792	CTBS-F/G	Cross-Validation
V	8.1	1000	CTBS-H + PRI-E	Equating
VI	9.1	1000	CTBS-J + PRI-E	Equating
VII	8.8	1000	CTBS-H/J	Equating
VIII	8.8	3935	CTBS-H/J	Cross-Validation

Table 2

The Eight Tests Studied and the Number of Items in Each[1]

Grade Combination	Test	Vocabulary Number of Items	Comprehension Number of Items
4/5	CTBS-F	23 (Test 1)	22 (Test 1)
	CTBS-G	22 (Test 2)	23 (Test 2)
	CTBS-F/G	45 (Test 1 + Test 2)	45 (Test 1 + Test 2)
	PRI-D	74 (Anchor Test)	
8/9	CTBS-H	23 (Test 1)	22 (Test 1)
	CTBS-J	22 (Test 2)	23 (Test 2)
	CTBS-H/J	45 (Test 1 + Test 2)	45 (Test 1 + Test 2)
	PRI-E	64 (Anchor Test)	

[1]All CTBS subtests contained a total of 45 items. However, only those items included in the interlevel linking test were used in this study.

Table 3

Number of Items, Raw score Mean and Standard Deviation, and Pearson Product-Moment Correlations by Sample

Vocabulary

Sample	Grade	n	Test 1 Test Level	Test 1 # Items	Test 1 $\bar{X}_1$	Test 1 SD_1	Anchor Test Test	Anchor Test # Items	Anchor Test $\bar{X}_a$	Anchor Test SD_a	Test 2 Test Level	Test 2 # Items	Test 2 $\bar{X}_2$	Test 2 SD_2	Correlations r_{1a}	Correlations r_{2a}	Correlations r_{12}
I	4.1	1000	F	23	13.90	6.00	PRI-D	74	45.38	14.09					.77		
II	5.1	1000					PRI-D	74	51.70	14.16	G	22	13.01	6.07		.75	
III	4.8	1000	F	23	15.54	6.02					G	22	12.26	6.14			.87
IV	4.8	1792	F	23	15.39	5.96					G	22	12.07	6.02			.86
V	8.1	1000	H	23	15.08	5.75	PRI-E	64	46.07	12.06					.71		
VI	9.1	1000					PRI-E	64	49.09	11.32	J	22	12.44	5.22		.66	
VII	8.8	1000	H	23	15.55	5.48					J	22	12.27	5.18			.83
VIII	8.8	3935	H	23	15.35	5.66					J	22	11.98	5.21			.84

Table 4

Number of items, raw score, mean and standard deviation, and Pearson product-moment correlations by sample.

Comprehension

Sample	Grade	n	Test 1 Test Level	Test 1 # Items	Test 1 $\bar{X}_1$	Test 1 SD_1	Anchor Test Test	Anchor Test # Items	Anchor Test $\bar{X}_a$	Anchor Test SD_a	Test 2 Test Level	Test 2 # Items	Test 2 $\bar{X}_2$	Test 2 SD_2	Correlations r_{1a}	Correlations r_{2a}	Correlations r_{12}
I	4.1	1000	F	22	13.50	5.94	PRI-D	74	45.38	14.09					.80		
II	5.1	1000					PRI-D	74	51.70	14.16	G	23	14.68	6.43		.69	
III	4.8	1000	F	22	13.61	5.76					G	23	13.45	6.27			.85
IV	4.8	1792	F	22	13.38	5.72					G	23	13.14	6.27			.87
V	8.1	1000	H	22	13.29	6.56	PRI-E	64	46.07	12.06					.61		
VI	9.1	1000					PRI-E	64	49.09	11.32	J	23	14.80	5.54		.71	
VII	8.8	1000	H	22	15.69	5.25					J	23	13.76	5.90			.81
VIII	8.8	3935	H	22	15.49	5.36					J	23	13.65	5.95			.82

Table 5

Mean and Standard Deviation of the True Score Estimates for Tests 1 and 2 Obtained for the Cross-Validation Sample, the Ratio of the Standard Deviations, the Scaled Root Mean Squared Difference, and Pearson Product-Moment Correlation

Vocabulary

Equating Method	Grade	n	Test 1				Test 2					RMSD	
			Test Level	# Items	$\bar{\hat{\zeta}}_1$	$SD_{\hat{\zeta}_1}$	Test Level	# Items	$\bar{\hat{\zeta}}_2$	$SD_{\hat{\zeta}_2}$	SD1/SD2	$S_{\hat{\zeta}}$	$r_{\hat{\zeta}_1\hat{\zeta}_2}$
1	4.8	1792	F	23	.626	.253	G	22	.622	.249	1.016	.478	.886
2					.630	.255			.616	.251	1.016	.481	.886
3					.622	.257			.619	.254	1.012	.475	.887
1	8.8	3935	H	23	.608	.229	J	22	.618	.230	.996	.512	.870
2					.597	.228			.626	.229	.996	.537	.864
3					.615	.238			.611	.232	1.026	.512	.869

Table 6

Mean and Standard Deviation of the True Score Estimates for Tests 1 and 2 Obtained for the Cross-Validation Sample, the Ratio of their Standard Deviations, the Scaled Root Mean Squared Difference, and Pearson Product-Moment Correlation

Comprehension

Equating Method	Grade	n	Test 1 Test Level	Test 1 # Items	Test 1 $\bar{\hat{\zeta}}_1$	Test 1 $SD_{\hat{\zeta}_1}$	Test 2 Test Level	Test 2 # Items	Test 2 $\bar{\hat{\zeta}}_2$	Test 2 $SD_{\hat{\zeta}_2}$	SD1/SD2	RMSD $S_{\hat{\zeta}}$	$r_{\hat{\zeta}_1\hat{\zeta}_2}$
1	4.8	1792	F	22	.597	.252	G	23	.608	.252	1.000	.505	.873
2					.599	.250			.607	.252	.992	.503	.874
3					.605	.255			.603	.251	1.016	.492	.879
1	8.8	3935	H	22	.684	.228	J	23	.618	.256	.890	.638	.839
2					.680	.226			.620	.258	.880	.629	.840
3					.650	.246			.650	.242	1.017	.550	.849

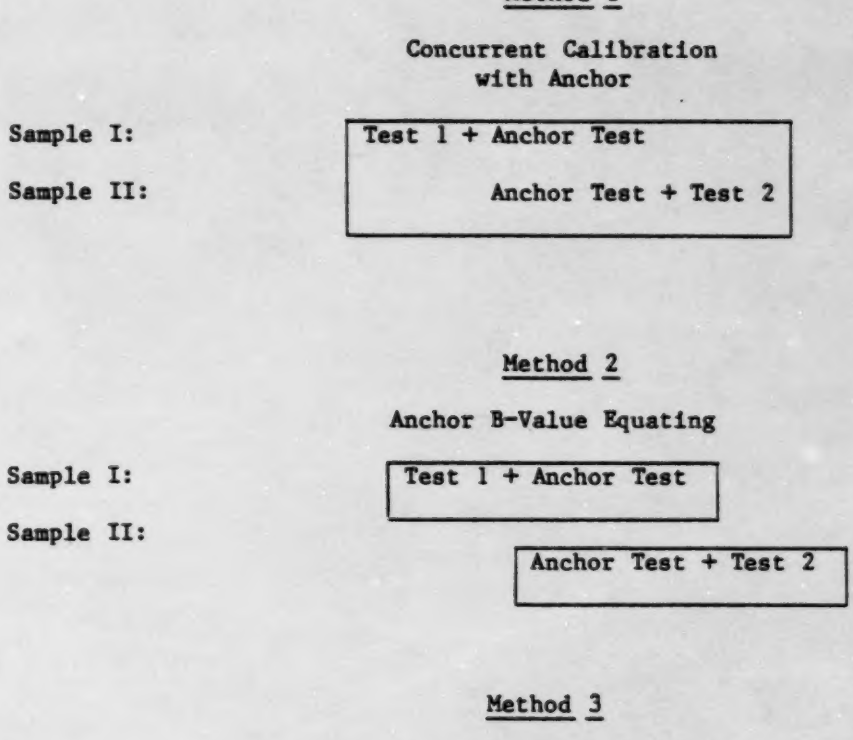

Figure 1. A schematic representation of the three vertical equating methods studied.